2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

An Online Automatic Calibration Method for Infrastructure-Based LiDAR-Camera via Cross-modal Object Matching

Tao Wang, Yuesheng He, Hanyang Zhuang, and Ming Yang

Abstract— **In indoor environments where the Global Navigation Satellite System (GNSS) isn't available, the infrastructure-based LiDAR-camera joint array can provide high-precision localization for mobile robots, such as Autonomous Valet Parking (AVP). The primary challenge in employing the infrastructure-based LiDAR-camera joint array is the extrinsic calibration between the LiDAR and the camera. Moreover, to handle interference deviation caused by vibrations or inadequate mounting stiffness during operation, the calibration's extrinsic parameters must be automatically updated online, presenting higher demands for infrastructure-based LiDAR-camera extrinsic calibration. This paper proposes an infrastructure LiDAR-camera online automatic calibration method based on prior knowledge of cross-modal target registration. This method requires no manual targets and initial pose guesses and can achieve extrinsic calibration. The object-prior model based on a lightweight object detection algorithm can rapidly detect scenes favorable for extrinsic calibration in sub-images of camera images. This creates favorable conditions for the registration of cross-modal networks and poses optimization of the LiDAR camera. Additionally, because a lightweight algorithm is used, the process does not compromise efficiency or consume excessive computational resources. Experimental results demonstrate that the proposed calibration method is suitable for calibrating infrastructure-based LiDAR-camera, with comparable accuracy and the ability to perform online calibration. Comparative experiments also show that the object-prior model can indeed select better scenes for LiDAR-camera extrinsic calibration, thus improving the accuracy and stability of extrinsic calibration to some extent.**

I. INTRODUCTION

Many indoor localization and navigation scenarios, such as Autonomous Valet Parking (AVP), require high-precision localization without relying on the Global Navigation Satellite System (GNSS), which is challenging. In indoor scenarios, achieving high-precision localization and navigation with a single intelligent agent requires the installation of complex and relatively expensive sensors on the mobile robot. Moreover, the cost increases rapidly with the number of intelligent agents. Infrastructure-based applications do not require extensive modifications on the object (such as vehicles), and the cost decreases as the number of localization and navigation objects increases. Compared to the former approach, infrastructure-based localization and navigation have become the ideal choice for indoor scenarios.

Tao Wang, Yuesheng He, and Ming Yang are with the Department of Automation, Shanghai Jiao Tong University, Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China (email: heyuesh@sjtu.edu.cn).

Hanyang Zhuang is with the University of Michigan - Shanghai Jiao Tong University Joint Institute. Shanghai Jiao Tong University, Shanghai, 200240, China.

Infrastructure-based Ultra-WideBand (UWB) networks are classic indoor high-precision localization and navigation solutions[1]. However, it relies heavily on onboard tags, making it unsuitable for scenarios with multiple agents and unable to perceive other obstacles. Localization based on infrastructure multi-camera networks[2], [3], [4], [5], [6], [7] suffers from lower accuracy due to incomplete object representation in images and challenges in directly estimating object poses from images, which is insufficient to meet the requirements of AVP (error less than 10 cm). The network based on environment-embedded LiDAR sensors demonstrated high accuracy in vehicle localization and tracking in parking garages[8]. However, it is prone to failure due to occlusions by buildings in the environment. Our previous work was based on the infrastructure with multi-RGB-D cameras network[9], achieving high accuracy in vehicle localization through adaptive modeling. However, RGB-D cameras have a small field of view (FOV), requiring many cameras for full coverage, leading to high cumulative costs and significantly limiting further large-scale deployment.

Summarizing the above analysis shows that infrastructure-based single-type sensor networks have significant limitations in application. Integrating LiDAR and cameras in the infrastructure of a LiDAR-camera joint array can effectively complement each other's shortcomings. This integration fully provides rich environmental texture information while accurately measuring the environment's geometric structure and three-dimensional features, which can effectively provide high-precision localization capabilities for mobile robots.

The fundamental task of integrating infrastructure-based LiDAR-camera arrays is the precise extrinsic calibration between the camera and LiDAR. Infrastructure-based LiDAR-camera arrays serve the indoor large-scale deployment market in different locations, making high automation necessary. To avoid increasing maintenance costs due to long-term or short-term offset disturbances such as vibration or insufficient installation rigidity, online extrinsic calibration for LiDAR-camera is a potential requirement. Thus, the potential challenge of integrating infrastructure-based LiDAR-camera arrays lies in the highly automated and real-time online extrinsic calibration between the camera and LiDAR.

Over the past two decades, research in LiDAR-camera extrinsic calibration has been actively pursued, including studies on finding 3D-2D feature correspondences of LiDAR-camera[10],[11],[12],[13] and then inputting these correspondences into various optimization algorithms for estimating LiDAR-camera relative pose. Some studies have designed specific artificial targets resembling chessboards to simplify

cross-modality image-to-lidar registration[11],[14]. In contrast, others have proposed calibration methods that do not require specific artificial targets[15],[16], estimating LiDAR-camera extrinsic parameters through matching low-level or semantic cross-modal features. These methods typically require a rough initial pose estimation as a prerequisite.

This paper proposes an online extrinsic calibration method for infrastructure-based LiDAR-camera based on cross-modal object matching. The method uses a lightweight object detection algorithm to select scenes favorable for extrinsic calibration. It utilizes a cross-modal registration network to register corresponding objects in the overlapping FOVs between the image and the range image of the point clouds. The contributions of this work are summarized as follows:

- We proposed a LiDAR-camera extrinsic calibration method via cross-modal object matching, achieving high automation without requiring initial pose estimation or a particular object.
- An object-prior model based on a lightweight object detection algorithm is proposed. The model rapidly detects scenes favorable for extrinsic calibration based on the sub-images of camera images, thereby creating favorable conditions for the registration of cross-modal networks and pose optimization of the LiDAR camera without reducing efficiency or consuming much computing power.

II. Related work

Over the past few decades, numerous automatic LiDAR-camera extrinsic calibration methods have been proposed. Significant progress has been made in methods requiring specific targets, as well as those not requiring specific targets, and whether rough initial pose estimation is needed.

A. Target-Based LiDAR-Camera Extrinsic Calibration

One of LiDAR-camera extrinsic calibration's most challenging aspects is finding correspondences between 3D and 2D features. A commonly used approach for LiDAR-camera extrinsic calibration is based on a specific target. The main idea is to design and create a robust and accurately detectable target that both LiDAR and the camera can simultaneously observe. By solving the perspective-n-point problem using the three-dimensional coordinates of points on the target and their two-dimensional projections in the image, the LiDAR-camera extrinsic is obtained. Particular geometric or intensity pattern targets have been designed to simplify the search for cross-modal feature correspondences, such as known geometric shapes for single target design[17], checkerboards[11],[14], or spherical targets[18]. These methods focus on finding correspondences between the three-dimensional coordinates and their two-dimensional projections on specific targets, making them suitable for vehicle-mounted LiDAR-camera extrinsic calibration and infrastructure-based LiDAR-camera extrinsic calibration.

B. Target-Free LiDAR-Camera Extrinsic Calibration

Another class of methods, different from target-based methods, directly extracts features from natural scenes, also known as target-free calibration methods[19]. Research on LiDAR-camera fusion algorithms for intelligent vehicle environment perception has matured. [20],[10],[21]propose fully automatic LiDAR-camera calibration without any prior information. [16] proposes online LiDAR-camera calibration given initial LiDAR-camera pose guesses. The fully automatic LiDAR-camera calibration method proposed by [19] requires no prior information and demonstrates good online performance. [22] proposes a generic LiDAR-camera calibration tool for LiDAR and camera projection models, suitable for more application scenarios, especially geometrically rich indoor scenes, but at the cost of poor online performance. [19] proposed an Attention-to-Optimization Approach for Automatic LiDAR-Camera Calibration via Cross-Modal Object Matching, which exhibits good online performance but suffers from poor stability due to the influence of online calibration scenes.

Our previous work implemented the calibration of a multi-RGB-D camera network using a chessboard board[23].

In contrast to the approaches above, our proposed calibration method does not require any artificial targets or initial pose guesses. It can be applied to calibrate the extrinsic parameters of infrastructure-based LiDAR-camera online, exhibiting high automation and stability.

III. METHOD

As shown in Fig.1, the online automatic calibration method of infrastructure-based LiDAR-camera consists of five steps. Before installation, we performed intrinsic calibration and distortion correction for the camera. After starting the experimental platform, we can obtain LiDAR point cloud frames and corresponding RGB images with time synchronization. Firstly, we enable a lightweight object detection algorithm to perform prior target screening, extracting scenes favorable for extrinsic calibration. Then, the filtered LiDAR-camera data pairs are inputted into a cross-modal network for object matching within the overlapping FOVs. Subsequently, 2D-3D feature data pairs are collected based on the object-matching relationships from the second step. Finally, point pair optimization and pose optimization are performed separately using fitness functions. Fig.2 illustrates our experimental platform, mainly consisting of LiDARs, cameras, servers, and accessories that together form a complementary intelligent infrastructure array of LiDARs and cameras, providing high-precision localization capabilities for indoor mobile robots. Compared to our previous study [9], the cost of a single camera is much lower than that of an RGB-D camera, and due to the larger FOV of the camera, the number of deployments is significantly reduced.

The method consists of four steps, each corresponding to a module:

- Object-Prior Model: This model selects scenes favorable for the calibration algorithm and excludes scenes with no objects or objects at the edge of the FOV.

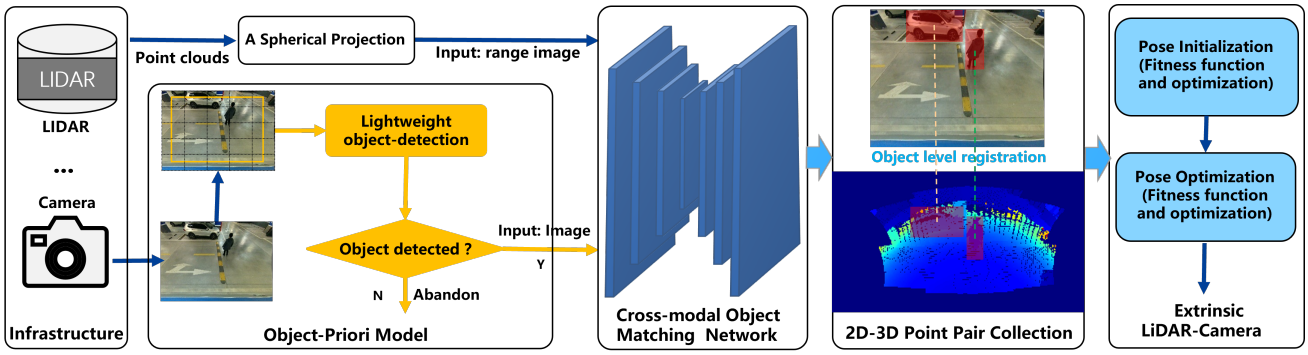


Fig. 1. Method overview. It uses a lightweight object detection algorithm to select scenes favorable for extrinsic calibration. Then, it uses a cross-modal registration network to register the corresponding objects in the overlapping FOVs between the image and the range image of point clouds. Based on the registered objects' correspondence, 2D-3D feature point pairs are collected, and the pose is initialized and optimized using fitness functions.

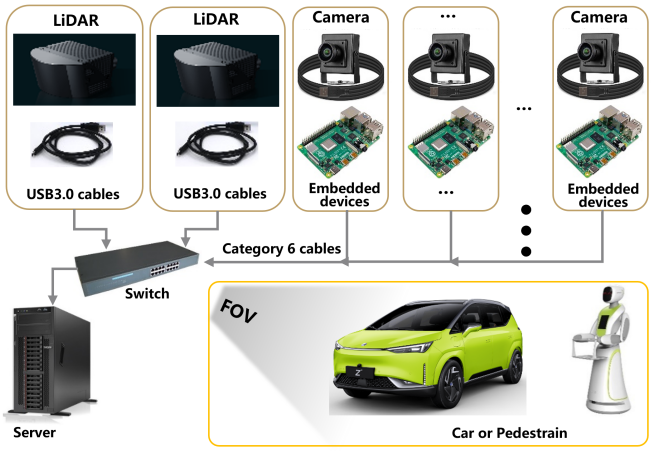


Fig. 2. Experimental Platform: Our platform consists of LiDAR, cameras, servers, and accessories that form a complementary intelligent infrastructure array of LiDAR and cameras.

- Cross-Modal Object Matching Module: This module registers LiDAR and camera data pairs based on objects in the overlapping FOVs, obtaining object-level correspondence.
- 2D-3D Point Pair Collection Module: This module collects 2D-3D feature point pairs based on cross-modal object correspondence.
- Position Initialization and Pose Optimization Module: This module iteratively refines 2D-3D point pairs using the Point-PSO algorithm[19] to solve for the initial extrinsic (R_0, T_0). and aligns object pixels with object point clouds using the Pose-PSO algorithm[19] to directly optimize the extrinsic (R, T).

A. Object-Prior Model

The object-prior model is an innovative design of our calibration method. The primary objective of extrinsic calibration is to obtain accurate extrinsic calibration results, rather than every set of LiDAR-camera FOV scenes being suitable for online calibration. Conducting online calibration using scenes with no targets or targets that are significantly incomplete in overlapping FOVs not only leads to computational waste but may also introduce substantial calibration errors, thus compromising the stability of the LiDAR-camera extrinsic.

The object-prior model is proposed to address this issue. The main idea of the object-prior model is to achieve lightweight and rapid object detection, identifying whether there are detected objects (pedestrians or cars) in sub-images of camera images. We utilize the lightweight Yolo-Fastest algorithm for its exceptional speed and lightweight computational power, enabling rapid completion of object priors with negligible time consumption and minimal strain on the computational resources of the on-site infrastructure system. Assuming the camera image is denoted as I with dimensions $W \times H$, and the sub-image of the camera image is denoted as I_{sub}, the object-prior model can be represented as follows:

$$O = \textbf{YOLO-Fastest}(I_{sub}) \tag{1}$$

In (1), YOLO-Fastest, the fastest and lightest version of the improved YOLO general object detection algorithm, has the object-prior model O output, which indicates whether the object (pedestrian or car) is detected in the sub-graph of the camera image.O is a binary value, where '1' represents the detection of least one object, and '0' represents no detection. I_{sub} represents the sub-image extracted from the camera image I, with dimensions $0.8W \times 0.8H$, centered at the original image, as shown in Fig.1.

B. Cross-Modal Object Matching Module

The structure of the cross-modal registration network is based on the CMON [19] network, with modifications. It mainly consists of two feature embedding modules: CNN and graph embedding. Assuming there are m objects O_I in the input image I and n objects O_R in the range image R (The range image is obtained by projecting the point cloud onto a spherical surface), the features of these objects are sampled from their centers and denoted as X_I and X_R, respectively. The CNN embedding module includes two parallel networks (E_I, E_R). E_I and E_R output feature maps (F_I, F_R) and the initial object features X_I^{ini} and X_R^{ini} are sampled through

bilinear interpolation:

$$F_I, F_R = \textbf{CNN}(I), \textbf{CNN}(R) \tag{2}$$

$$X_I^{ini} = \textbf{Iterp}(O_I, F_I) \qquad X_I^{ini} \in R^{m \times d} \tag{3}$$

$$X_R^{ini} = \textbf{Iterp}(O_R, F_R) \qquad X_R^{ini} \in n^{n \times d} \tag{4}$$

The graph embedding module aims to capture more context information to enhance the quality of object features. It comprises five stacked multi-head self-attention modules (MHSA)[24]. ζ is the expression of MHSA, and the graph embedding is represented as:

$$X_I = \zeta_5(\cdots(\zeta_1(X_I^{ini}))) \tag{5}$$

$$X_R = \zeta_5(\cdots(\zeta_1(X_R^{ini}))) \tag{6}$$

Due to the small number of objects in the overlapping FOVs, it is not possible to construct an effective graph structure. We generate auxiliary grid points (Z_I, Z_R) in the feature map to capture denser context information, as shown in Fig.3. These points are all sent into the graph embedding module for extracting context information, and the graph embedding is represented as:

$$(\textbf{cat}\{X_I, Z_I\}) = \zeta_5(\cdots(\zeta_1(\textbf{cat}\{X_I^{ini}, Z_I^{ini}\}))) \tag{7}$$

$$(\textbf{cat}\{X_R, Z_R\}) = \zeta_5(\cdots(\zeta_1(\textbf{cat}\{X_R^{ini}, Z_R^{ini}\}))) \tag{8}$$

The "cat{,}" denotes feature concatenation. Using the obtained object features (X_I, X_R), we can calculate the metric function M to evaluate object-level similarity:

$$M = \textbf{sig}\big((X_I)^T X_R\big) \tag{9}$$

In (9), the "sig" represents the sigmoid function where we use inner production to calculate the feature similarity.

$$P(O_R^{j_0} \in FOV_I) = p(j_0) = \textbf{Interp}(O_R^{j_0}, \chi) \tag{10}$$

Where $p(j_0)$ represents the probability that the $j_0 - th$ 3D object is in the FOV of the image I, obtained through bilinear interpolation sampling. The FOV-attention map is denoted as χ. We describe the object-matching probability in the form of a joint distribution:

$$S(O_I, O_R) = \textbf{Softmax}(M(O_I, O_R) \cdot P(O_R \in FOV_I)) \tag{11}$$

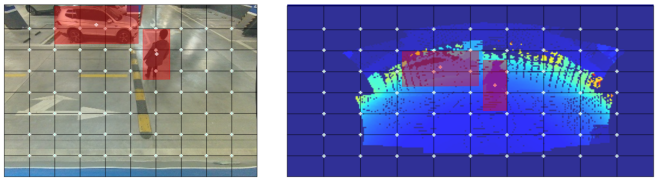
Input synchronized LiDAR-camera images: first, convert the point cloud into a range image, then input it into a trained network(CMON) to obtain target correspondence in the overlapping FOVs, as shown in (12).

$$(\textbf{cat}\{X_I, Z_I\}), (\textbf{cat}\{X_R, Z_R\}) = \textbf{CMON}(I, R) \tag{12}$$

C. 2D-3D Point Pair Collection Module

Based on the cross-modal object correspondence, collect the center point and multiple feature points of the matched objects, obtaining 2D-3D point pairs $(p_I^i, p_R^i) | p_R = (x, y, z), p_I = (u, v), i = 1, ..., N$. Then, the rotation and translation matrix (R, T) are solved using the EPnP[25] algorithm.

$$(u_i, v_i) \leftrightarrow (X_i, Y_i, Z_i) \tag{13}$$



(a) (b)

Fig. 3. Feature Sampling. Besides the points provided by the object centers, we generate grid points in the image and range image plane to capture more dense contextual information. All of these points are fed into the graph embedding module to extract structure information.

D. Position Initialization and Pose Optimization Module

Pose initialization and pose optimization are optimized twice using the fitness function(14). First, because it is difficult to guarantee the accuracy of the 2D-3D point pairs collected in the third step, we use the alignment function (14) to evaluate the alignment between image pixels and point clouds, gradually refining the 2D-3D point pairs to converge to the correct positions as much as possible. Then, the refined 2D-3D pairs are used to solve the initial extrinsic (R_0, T_0) using the RANSAC-EPnP[25] algorithm.

$$F(\Phi) = \sum_{i=1}^{n} Q(u_i, v_i | p_i, \Phi) \tag{14}$$

where n is the number of the object point clouds, the image coordinates (u_i, v_i) of the 3D points pi are calculated by using the estimated extrinsic parameters of particle Φ. Next, based on the solved initial pose, the alignment function (1) is used again to align object pixels with object point clouds, directly optimizing to obtain (R, T).

IV. EXPERIMENT

A. Experiment Setup

We equipped the indoor infrastructure-based LiDAR-camera, as illustrated in Fig.4, to capture two sets of LiDAR-camera relative poses. We collected scenes with "pedestrian" and "car" appearances in the overlapping FOV of each LiDAR-camera pair online in Fig.5.

To compare the system's performance systematically, we used the algorithm proposed by Jiunn-Kai Huang et al.[26] and the classical chessboard-based pose estimation algorithm[27] for ground truth calibration of the external parameters of the infrastructure laser scanner-camera setup.

In our article, we trained on a limited number of objects. The training data includes data generated by combining with camera networks[9] and open-source KITTI odometry datasets(sequences 0 to 8)[28]. The average pose error (APE) is also used as a quantitative evaluation metric for the experiments (APE measured in meters and degrees). The LiDAR and camera are connected to the computer via Ethernet cables, and all experiments are conducted on a computer running Ubuntu 18.04 LTS (64-bit) with an AMD R9-7950HX and RTX 3090.

Fig. 4. Scene Installation Diagram. Two sets of infrastructure LiDAR-camera pairs are installed at positions A and B, respectively, and positions A and B are located on both sides of the parking lot road and towards the parking lot ground.

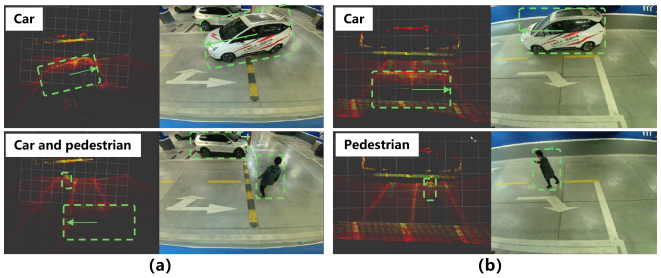


Fig. 5. Datasets. We continuously collected point cloud data and image information from LiDAR-camera pairs at positions A and B over some time. (a) and (b) are the scenes featuring "pedestrian" and "car" from the overlapping FOVs of the LiDAR-camera pairs at position A and B, respectively.

B. Experiments on External Parameter Calibration

We extracted 13400 synchronized LiDAR-camera data pairs from two sets of independent LiDAR-camera collected scenes. After filtering the camera's original images using Yolo-Fastest, we obtained 2887 data pairs containing detected 'pedestrian' or 'car' objects, with 1874 from position A and 1013 from position B. Additionally, after filtering the sub-images of the camera's original images, we obtained 1084 synchronized LiDAR-camera data pairs, with 708 from position A and 376 from position B. We used the data after object-prior detection for cross-modal registration, point pair extraction, pose initialization, and pose optimization.

To compare the performance of the proposed method, we calibrated the extrinsic parameters of two LiDAR-camera pairs using the calibration algorithm proposed by Koide et al. [22] and the ATOP method proposed by Sun et al. [19]. and then calculated the calibration errors for both methods.

Table.I summarizes the calibration accuracy errors using different methods and provides statistics on the calibration efficiency of our proposed method. As expected, our method filters the calibration scene data pairs and achieves minor errors at both positions. Compared with the method proposed by Koide, Kenji et al., our method does not require accumulating point clouds over time, making it more suitable for online calibration. Our calibration method uses better and fewer synchronized LiDAR-camera data pairs than the ATOP method. Our method achieves better calibration accuracy in the same time sequence with fewer synchronized LiDAR-camera data pairs, consuming roughly the average time cost and saving overall system computational power.

TABLE I
EXTRINSIC CALIBRATION ERROR

Seq.	Position A			Position B		
	Rot.[°]	Trans.[m]	Time.[s]	Rot.[°]	Trans.[m]	Time.[s]
Koide, et al.[22]	0.273	0.085	–	0.198	0.069	–
ATOP[19]	0.178	0.051	<18	0.154	0.055	<15
Ours	**0.085**	**0.031**	<15	**0.093**	**0.029**	<15

C. Ablation Study

We conducted ablation studies on two sets of LiDAR-camera combinations to examine the overall impact of object prior detection on our method. All other modules remained unchanged. We divided the LiDAR-camera data pairs into four Combinations:

- Combination(1): Without any object prior detection.
- Combination(2): After object prior detection, filtering the camera's original image.
- Combination(3)-1: After object prior detection, filtering a sub-image of the camera's image, specifically a center sub-image of 4/5 of the length and width.
- Combination(3)-2: After object prior detection, filtering a sub-image of the camera's image, specifically a center sub-image of 3/5 of the length and width.

TABLE II
ABLATION STUDY

Experiments Method + Combination	Position A		Position B	
	Rot. [°]	Trans. [m]	Rot. [°]	Trans. [m]
ATOP[19] +Combination(1)	0.178	0.051	0.154	0.055
Ours+Combination(2)	0.173	0.050	0.182	0.047
Ours+Combination(3)-1	0.085	0.031	**0.093**	**0.029**
Ours+Combination(3)-2	**0.079**	**0.029**	0.105	0.031

Table.II summarizes the experimental results of the four combinations. Combination(1), without any filtering, introduced many synchronized LiDAR-camera data pairs with no objects or objects at the edge of the FOVs, resulting in a sizeable average calibration error. Although combination(2) filtered out LiDAR-camera data pairs without objects, those with edge objects still caused a significant calibration error. For combination(3), the two experiments successfully filtered out LiDAR-camera data pairs without objects or with edge objects, ensuring a higher-quality calibration scene and achieving a minor calibration error. We also found that combinations(3)-1 and (3)-2 achieved similar errors,

indicating that the choice of sub-image size from the camera's original image is not necessarily more accurate with a smaller sub-image. The smaller the sub-image selected and filtered, the fewer synchronized LiDAR-camera data pairs remained, leading to a different optimal ratio for each data combination, which requires specific data analysis. The experimental results demonstrate the importance of the object prior detection module in the proposed external parameter calibration method.

V. CONCLUSIONS

This paper proposes an online automatic calibration method for infrastructure LiDAR-camera based on object priors via cross-modal object matching. This method does not require any particular target or initial pose, enabling online extrinsic calibration with high automation and comparable accuracy. Additionally, the object priors model can filter out better scenes for LiDAR-camera extrinsic calibration, ensuring extrinsic calibration accuracy and stability to a certain extent.

ACKNOWLEDGMENT

We thank my mentor, Professor YueSheng He, for his guidance. We would like also to appreciate the insightful advice from Dr. HanYang Zhuang and Fei Wang. Additionally, we are thankful for the support from Dr. Yi Sun and Professor Jian Li from the National University of Defense Technology. This work was supported by the National Natural Science Foundation of China(U22A20100/62203294/62373250).

REFERENCES

[1] C. Zhang, M. Kuhn, B. Merkl, A. E. Fathy, and M. Mahfouz, "Accurate uwb indoor localization system utilizing time difference of arrival approach," in *2006 IEEE radio and wireless symposium*. IEEE, 2006, pp. 515–518.

[2] A. Ibisch, S. Houben, M. Michael, R. Kesten, and F. Schuller, "Arbitrary object localization and tracking via multiple-camera surveillance system embedded in a parking garage," in *Video surveillance and transportation imaging applications 2015*, vol. 9407. SPIE, 2015, pp. 130–141.

[3] J. Einsiedler, D. Becker, and I. Radusch, "External visual positioning system for enclosed carparks," in *2014 11th Workshop on Positioning, Navigation and Communication (WPNC)*. IEEE, 2014, pp. 1–6.

[4] J. Einsiedler, O. Sawade, B. Schäufele, M. Witzke, and I. Radusch, "Indoor micro navigation utilizing local infrastructure-based positioning," in *2012 IEEE Intelligent Vehicles Symposium*. IEEE, 2012, pp. 993–998.

[5] D. Becker, J. Einsiedler, B. Schäufele, A. Binder, and I. Radusch, "Identification of vehicle tracks and association to wireless endpoints by multiple sensor modalities," in *International Conference on Indoor Positioning and Indoor Navigation*. IEEE, 2013, pp. 1–10.

[6] A. Ibisch, S. Houben, M. Schlipsing, R. Kesten, P. Reimche, F. Schuller, and H. Altinger, "Towards highly automated driving in a parking garage: General object localization and tracking using an environment-embedded camera system," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 2014, pp. 426–431.

[7] T. Partanen, P. Müller, J. Collin, and J. Björklund, "Implementation and accuracy evaluation of fixed camera-based object positioning system employing cnn-detector," in *2021 9th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2021, pp. 1–6.

[8] A. Ibisch, S. Stümper, H. Altinger, M. Neuhausen, M. Tschentscher, M. Schlipsing, J. Salinen, and A. Knoll, "Towards autonomous driving in a parking garage: Vehicle localization and tracking using environment-embedded lidar sensors," in *2013 IEEE intelligent vehicles symposium (IV)*. IEEE, 2013, pp. 829–834.

[9] B. Cao, Y. He, H. Zhuang, and M. Yang, "Infrastructure-based vehicle localization system for indoor parking lots using rgb-d cameras," *Journal of Shanghai Jiaotong University (Science)*, vol. 28, no. 1, pp. 61–69, 2023.

[10] M. Feng, S. Hu, M. H. Ang, and G. H. Lee, "2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4790–4796.

[11] S. Verma, J. S. Berrio, S. Worrall, and E. Nebot, "Automatic extrinsic calibration between a camera and a 3d lidar using 3d point and plane correspondences," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 3906–3912.

[12] Z. Taylor, J. Nieto, and D. Johnson, "Multi-modal sensor calibration using a gradient orientation measure," *Journal of Field Robotics*, vol. 32, no. 5, pp. 675–695, 2015.

[13] L. Zhou, Z. Li, and M. Kaess, "Automatic extrinsic calibration of a camera and a 3d lidar using line and plane correspondences," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5562–5569.

[14] W. Wang, K. Sakurada, and N. Kawaguchi, "Reflectance intensity assisted automatic and accurate extrinsic calibration of 3d lidar and panoramic camera using a printed chessboard," *Remote Sensing*, vol. 9, no. 8, p. 851, 2017.

[15] J. Kang and N. L. Doh, "Automatic targetless camera–lidar calibration by aligning edge with gaussian mixture model," *Journal of Field Robotics*, vol. 37, no. 1, pp. 158–179, 2020.

[16] Y. Zhu, C. Li, and Y. Zhang, "Online camera-lidar calibration with sensor semantic information," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4970–4976.

[17] J. Domhof, J. F. Kooij, and D. M. Gavrila, "A joint extrinsic calibration tool for radar, camera and lidar," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 571–582, 2021.

[18] T. Tóth, Z. Pusztai, and L. Hajder, "Automatic lidar-camera calibration of extrinsic parameters using a spherical target," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8580–8586.

[19] Y. Sun, J. Li, Y. Wang, X. Xu, X. Yang, and Z. Sun, "Atop: An attention-to-optimization approach for automatic lidar-camera calibration via cross-modal object matching," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 696–708, 2022.

[20] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation using uncalibrated lidar and stereo fusion," *Ieee transactions on intelligent transportation systems*, vol. 21, no. 1, pp. 321–335, 2019.

[21] G. Iyer, R. K. Ram, J. K. Murthy, and K. M. Krishna, "Calibnet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1110–1117.

[22] K. Koide, S. Oishi, M. Yokozuka, and A. Banno, "General, single-shot, target-less, and automatic lidar-camera extrinsic calibration toolbox," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 301–11 307.

[23] H. Yuesheng, W. Tao, C. Long, Z. Hanyang, and Y. Ming, "An extrinsic calibration method for multiple infrastructure rgb-d camera networks with small fov," *IEEE Open Journal of Intelligent Transportation Systems*, pp. 1–1, 2024.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[25] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: Efficient perspective-n-point camera pose estimation," *Int. J. Comput. Vis*, vol. 81, no. 2, pp. 155–166, 2009.

[26] J.-K. Huang and J. W. Grizzle, "Improvements to target-based 3d lidar to camera calibration," *IEEE Access*, vol. 8, pp. 134 101–134 110, 2020.

[27] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[28] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.